UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

May 22, 2012

DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-08

MEMORANDUM FOR David C. Whitford
Chief, Decennial Statistical Studies Division

From: Patrick J. Cantwell *(Signed)*
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Colt Viehdorfer
Mark Seiss
Gregg Diffendal
Decennial Statistical Studies Division

Subject: 2010 Census Coverage Measurement Estimation Report: Missing Data for Net Coverage Estimation

This report is one of twelve documents providing estimation results from the 2010 Census Coverage Measurement program. This report focuses on missing data for net coverage estimation for persons and housing units for the United States - inclusion and match status for the P sample, and enumeration status for the E sample.

For more information, contact Colt Viehdorfer on (301) 763-6796 or Mark Seiss on (301) 763-9283.

cc:
DSSD CCM Contacts List

Census Coverage Measurement Estimation Report

Missing Data for Net Coverage Estimation

Prepared by
Colt Viehdorfer
Mark Seiss
Gregg Diffendal

Decennial Statistical Studies Division

Table of Contents

Executive Summary

The levels of missing data in the 2010 Census Coverage Measurement program were low. Further, the occurrence of unresolved status was not unduly clustered in any sample block cluster. Thus, the missing data procedures should have only a minor effect on the estimation.

- *Noninterviews in the P Sample.* The interview rate in the 2010 Census Coverage Measurement program was 81.7%, with a noninterview rate of 3.1%. Among occupied housing units, the interview rate was 96.4%. Due to the high response rate, most of the noninterview adjustment factors applied were very close to 1.

- *Unresolved **inclusion** status in the P Sample.* The proportion of people with unresolved inclusion status was 2.9%. Thus, it appears that missing this item has only a minor effect on the estimation process. The average rate of inclusion assigned to people with unresolved status was 0.68.

- *Unresolved **match** status in the P Sample.* Only 3.7% of the P-sample persons had unresolved match status. We assigned an average match rate of 0.62 to people with unresolved match status. The low rate of unresolved match status implies there is a small effect on the estimation.

- *Unresolved **enumeration** status in the E Sample.* About 4.8% of the people in the E sample had unresolved enumeration status. The average rate of correct enumeration assigned to people with unresolved status was 0.83.

- For housing units, the 2010 Census Coverage Measurement had very low missing data rates, less than 1.0% for all housing unit statuses.

1. Introduction

Before calculating dual system estimates, we had to account for missing information from the interviews of P-sample people and from the matching and followup operations. Note that the term "missing data" applied after all followup attempts were completed. We encountered three types of missing data in the Census Coverage Measurement (CCM) and used three procedures to correct for them.

1. *Household-level noninterviews in the P Sample.* In a majority of these, the household was unable to be contacted or the interview was refused. In general, the noninterview adjustment spread the weights of household noninterviews among households that were interviewed in the same block cluster (the primary sampling unit) and type of basic address (single family, multi-unit address, or other).

2. *Missing demographic characteristics.* This situation occurred when a person was missing age, sex, relationship, tenure, race, or Hispanic origin. When a housing unit was missing characteristics, we imputed tenure and race of the head of household for housing unit estimation. Results of characteristic imputation are located in Shores and Sands (2012).

3. *Unresolved status.* For some respondents in the P sample, there was not enough information available to determine the inclusion status (whether or not the person should have been included in the P sample), the mover status (whether or not the person was an inmover), or the match status (whether or not the person matched to someone enumerated in the census in the same block cluster or the extended search area). The residence status code was used to assign inclusion status. For housing units, unit status determined whether the housing unit was in the P sample or not. Match status could also be missing for housing units.

   Similarly, for people and housing units in the E sample, there may not have been enough information to determine whether the person or housing unit was correctly enumerated, resulting in unresolved enumeration status. Generally, for cases with missing status, a probability was assigned based on information available about the specific case and about resolved cases with similar characteristics.

Note that E-sample people without sufficient information (a name and at least two other characteristics) for matching were not unresolved for net coverage estimation, but were treated as erroneous enumerations, that is, they were assigned a probability of correct enumeration of 0. In the P sample, if the entire housing unit contained people without sufficient information for matching, the housing unit was treated as a noninterview. Otherwise, each such person in an occupied housing unit had an unresolved match status.

After applying methods to account for the three types of missing data, a weight trimming procedure was implemented prior to dual system estimation calculation to reduce the effect of block clusters that might have an undue effect on the estimates. Clusters were identified as being influential clusters if they had a large difference between the number of E-sample erroneous enumerations and P-sample nonmatches.

2. Methods

The 2010 CCM program measured the net coverage and components of census coverage in the 2010 Census and provided estimates for various demographic groups and levels of geography. In many ways, the CCM resembled the 2000 Accuracy and Coverage Evaluation (A.C.E.). The A.C.E. also suffered from missing data in terms of the components just described.

In the 2010 CCM, the Census Bureau applied statistical procedures to account for missing data similar to those used in the A.C.E. The 2010 method used to adjust for noninterviews in the P sample followed that used in the 1990 and 2000 coverage surveys. We spread the noninterviewed housing unit weights to interviewed housing unit weights within cluster and structure type, with collapsing if too many noninterviews were observed. We made one change for 2010 to account for a large number of noninterviews in some American Indian blocks. We changed the collapsing hierarchy to spread the weights for the noninterviewed units within American Indian blocks to other American Indian blocks in the same state.

In the 2010 CCM, we used a different mover procedure. Inclusion status (or residence status, in 2000) determined whether the person was excluded from the P sample. For the 2010 CCM, persons were processed so that nonmovers, inmovers, and some outmovers (those who had no chance of being captured in the P sample, e.g., people currently living in a group quarters facility or those who moved out of the country) were in the P sample. Others were not included in the P sample (never resident, outmovers who could be captured at their outmover address, and persons out of scope). If outmover persons were not determined to be in or out of the P sample, they were treated as being out of the P sample for estimation purposes.

The main difference between the missing data procedures for the 2000 A.C.E. and 2010 CCM dealt with the imputation of unresolved statuses. For both surveys, each person in the P sample had a probability of matching to a person in the census. This probability was said to be 1 if the person matched and 0 if the person did not match. People whose match status was "unresolved"—still unknown or unclear after all followup operations—were assigned a match probability between 0 and 1 to compute the dual system estimate (DSE). Analogous situations describe inclusion status for P-sample people and enumeration status for E-sample people in the 2000 A.C.E. and 2010 CCM.

In the 2000 A.C.E. procedure, all resolved and unresolved cases were separated into groups called imputation cells according to a set of operational and demographic characteristics. Within any cell, the weighted proportion of matches (or residents, or correct enumerations) among the resolved cases was assigned as the probability of a match to all unresolved cases in that cell. In the 2010 CCM procedure, all resolved cases were used in a logistic regression model to predict a probability for the unresolved cases. Separate logistic regression models were used to predict the P-sample match and inclusion statuses for cases with sufficient and insufficient information for matching.

3. Limitations

All of the missing data models assumed ignorability (Rubin 1976), which is that the probabilities of match, residence, and enumeration status given a set of known covariates are the same for resolved and unresolved cases.

For 2010, we used a post-enumeration survey (PES) B+ procedure, which calculated the match rate using nonmovers, inmovers, and outmovers who lived out of the country or in a group quarters on interview day. This differs from the 2000 procedure that used a PES C methodology. Under the PES C procedure, the match rate was computed using nonmovers and outmovers, while inmovers were used to estimate the number of movers. Since the match rates were calculated using different procedures, a direct comparison of the two rates has limitations.

In 2010, the person interview and followup interview occurred later in the year than in 2000. In 2010, the person interview was conducted from August through October, while the person followup occurred from January to March of the following year. In 2000, person interviewing took place from May to August, and the followup interview occurred from October to November. Conducting person interviews further from Census Day in 2010 CCM meant respondents might have been less knowledgeable or less cooperative. Different proxy rules and changes to the definition of a complete person interview may also affect the 2000 and 2010 comparison. See Linse and Stone (2010) for more details.

Most comparisons of the 2010 missing data rates with 2000 are to the original A.C.E. results. The Census Bureau subsequently issued the A.C.E. Revision II estimates that made adjustments for the overestimation of correct enumerations and matches. Analogous missing data rates were not computed, except for noninterview adjustment and weight trimming. It is important to note that the total estimate of correct enumerations decreased from 252.1 million in the 2000 A.C.E. to 248.3 million in the 2000 A.C.E. Revision II. P-sample results were also impacted but not to the same extent as E-sample enumeration status. See Haines and Mule (2003) or Mule (2001) for more information.

4. Discussion of Results

4.1 Noninterview Rates

Table 1 contains the summary of the person interview for the 2010 CCM and results from the 2000 A.C.E. Revision II. Vacant and deleted housing units were not used in the noninterview adjustment procedure. Only interviewed and noninterviewed housing units were used in the procedure and in calculating the interview rate for occupied housing units (96.4%). Comparisons with 2000 A.C.E. Revision II are difficult because two noninterview adjustment procedures were used, one for Interview Day and another for Census Day. The PES B+ procedure used in 2010 CCM is more comparable to the 2000 A.C.E. Revision II Interview Day procedure. However, the Census Day interviews were the major focus of the 2000 A.C.E. Revision II interview. Comparisons and explanations of differences in noninterview rates from the 2000 A.C.E. and 2010 CCM programs are located in Linse and Stone (2010).

Table 1. Summary of the 2010 CCM Person Interview

	2010 CCM		2000 A.C.E. Rev. II Interview Day[1]	2000 A.C.E. Rev. II Census Day[2]
	Count	Percent	Percent	Percent
Total Housing Units	171,217	100.0	100.0	100.0
Interview	139,956	81.7	89.1	85.6
Noninterview	5,285	3.1	0.9	2.8
Vacant	21,296	12.4	8.8	8.9
Deletes	4,680	2.7	1.1	2.8

[1] Results come from Table 1c in Ikeda and Beaghen (2002).
[2] Results come from Table 1a in Ikeda and Beaghen (2002).

4.2 Missing Data Results for Persons

Table 2 provides a summary of the different statuses, P-sample inclusion and match by sufficient and insufficient information, and E sample enumeration status, with the number unresolved and the mean imputed value. Sufficient information cases were cases with a name and at least two characteristics, while insufficient information cases had no name or fewer than two characteristics. All insufficient information cases were unresolved for inclusion and match statuses.[1] All mean imputed values in section 4.2 are weighted means.

Table 2. 2010 CCM Imputation of Statuses

	Number unresolved	Mean Imputed Value
P-Sample Inclusion Status	11,206	0.68
Sufficient Information	4,792	0.76
Insufficient Information	6,414	0.61
P-sample Match Status	13,002	0.62
Sufficient Information	6,588	0.45
Insufficient Information	6,414	0.85
E-Sample Enumeration Status	18,522	0.83

Note: The P-sample total is 392,711 records and the E-sample total is 383,537 records.

4.2.1 Missing Inclusion Status

Table 3 contains a summary of the inclusion status results, separately for sufficient and insufficient information cases. Inclusion status determines whether a case should be included in the P sample or not. Unresolved cases had their weights adjusted down by the probability of being included in the P sample. Thus, all of the unresolved cases were included but were downweighted.

[1] Cases in vacant or deleted housing units with insufficient information for matching were excluded from the P sample.

Table 3. 2010 CCM Inclusion Status

	Total People	Percent	Sufficient Information for Matching	Insufficient Information for Matching
U.S total	392,711	100.0	379,679	13,032
In P sample	355,812	90.6	349,398	6,414
Resolved	344,606	87.8	344,606	0
Unresolved	11,206	2.9	4,792	6,414
Not In P sample	36,899	9.4	30,281	6,618

4.2.2 Missing Match Status

Table 4 contains a summary of the match status for the P-sample persons, separately for mover status. There are sizeable differences for nonmovers, inmovers, and outmovers. Most people with an unresolved match status were inmovers since we needed to know their exact Census Day address and had to be able to geocode that address in order to search for the person to call them a match or nonmatch. There were almost no unresolved match statuses for nonmovers and outmovers since their census day address is their interview day address. Note that all unresolved inclusion status cases were also unresolved for mover and match status.

Table 4. 2010 CCM Match Status by Mover Status

P Sample	Total People	Match Rate	Nonmatch Rate	Unresolved Match Rate	Mean Imputed Value
2010 CCM	355,812	87.7%	8.7%	3.7%	0.62
Nonmover	316,977	92.1%	7.9%	0.0%	0.55
Inmover	26,399	71.5%	21.8%	6.7%	0.67
Outmover	1,230	84.3%	15.7%	0.0%	NA
Unresolved Mover	11,206	0	0	100%	0.61
2000 A.C.E.[1]	640,945	90.3%	8.5%	1.2%	0.84
Nonmover	617,490	91.1%	8.0%	0.9%	NA
Outmover	23,455	67.8%	21.7%	10.5%	NA

[1] Missing data results from the 2000 A.C.E. can be found in Cantwell et al. (2001).
Note: The 2000 A.C.E. does not include inmovers since it used the PES-C procedure.

4.2.3 Missing Enumeration Status

Table 5 contains a summary of the enumeration status for E-sample persons. Note that all E-sample cases with insufficient information were considered resolved as erroneous enumerations for net coverage estimation. Census whole-person imputation cases were not considered in the E sample and are not included in Table 5.

Table 5. 2010 CCM and 2000 A.C.E. Enumeration Status

E Sample	Total People	Correct Enumeration Rate	Erroneous Enumeration Rate	Unresolved Enumeration Rate	Mean Imputed Value
2010 CCM	383,537	87.3%	7.8%	4.8%	0.83
2000 A.C.E.[1]	704,602	92.6%	4.4%	3.0%	0.95

[1] Missing data results from the 2000 A.C.E. can be found in Cantwell et al. (2001).

4.3 Missing Data Results for Housing Units

For housing units, a status was easier to resolve since there were no movers and the concept of sufficient or insufficient information for matching was not applicable. The number of unresolved cases was very small. The impact on the estimates due to missing data should be very small. In the 2010 CCM program, there are 166,877 P-sample housing units and 172,503 E-sample housing units. All mean imputed values in section 4.3 are weighted means.

Table 6. Imputation of Statuses - Housing units

Status	Number Unresolved	Percent Unresolved	Mean Imputed Value
2010 CCM			
P-Sample HU	58	0.0%	0.86
P-Sample Match	10	0.0%	0.84
E-Sample Enumeration	215	0.1%	0.94
2000 A.C.E.[1]			
P-sample HU	31	0.0%	NA
P-sample Match	25	0.0%	NA
E-sample Enumeration	688	0.2%	NA

[1] Housing Unit missing data results from the 2000 A.C.E. can be found in Chen et al. (2007).
Note: The 2000 A.C.E. P sample had 300,913 HUs and the E sample had 311,029 HUs.

4.3.1 Missing Housing Unit Status

The status of a housing unit refers to whether the listed housing unit should be in the P sample or not. The listing included addresses that were not housing units at the time of the listing, but might have become housing units by the time of CCM interviewing. The number and percent of cases that were unresolved was very small. The mean imputed probability of being a HU for the 58 unresolved cases was 0.86.

Table 7. 2010 CCM Housing Unit Status

	Total Housing Units	Percent
2010 CCM Independent sample[1]	171,217	100.0%
Resolved – In sample	166,819	97.4%
Resolved – Not In sample	4,340	2.5%
Unresolved	58	0.0%

[1] Independent sample includes the HUs available for CCM person interviewing after the subsampling is completed. Census only HUs are not included.

4.3.2 Missing Housing Unit Match Status

Only 10 housing unit records had an unresolved match status. The housing unit match rate for 2010 CCM was high with over 96% of the housing units matched to a census unit. The mean imputed match rate for the 10 unresolved records was 0.84.

Table 8. 2010 CCM Housing Unit Match Status

	Total Housing Units	Percent
In P sample	166,877	100.0%
Matched	160,644	96.3%
Not Matched	6,223	3.7%
Unresolved	10	0.0%

Note: In the 2000 A.C.E., 90.83% of P sample records were matches. More information is located in Chen et al. (2007).

4.3.3 Missing Housing Unit Enumeration Status

Missing housing unit enumeration status was very low, under 0.2%. Unresolved enumeration status should have almost no effect on the housing unit estimation. The mean imputed correct enumeration rate for the 215 unresolved records was 0.94.

Table 9. 2010 CCM Housing Unit Enumeration Status

	Total Housing Units	Percent
In E sample	172,503	100.0%
Correct Enumeration	166,449	96.5%
Erroneous Enumeration	5,839	3.4%
Unresolved	215	0.1%

Note: In the 2000 A.C.E., 95.8% of E sample records were correct enumerations. More information is located in Chen et al. (2007).

4.4 Weight Trimming

In 2010 CCM, few clusters required weight trimming. For persons, only one cluster had an undue influence on the net coverage estimates. The outlier cluster was an American Indian Reservation (AIR) cluster. Person weights within the cluster were trimmed to 78.5% of the original weight. The cluster did not have any E-sample persons.

In 2010 CCM, two clusters required weight trimming for housing units. The first cluster was a non-AIR cluster. This cluster was determined to be an outlier because of a large number of housing units that were erroneous enumerations because they were determined to be group quarters. The final weights of the housing units in this cluster were 76.5% of the original weight. The second outlying housing unit cluster was the same AIR cluster that required trimming for persons. There were not any E-sample housing units in this cluster, but the P-sample weights were trimmed to 43.0% of their initial value.

In 2000, one cluster was trimmed for persons. It was a non-AIR cluster. For more details on 2000 A.C.E. person weight trimming results see Mule (2003). A total of six clusters were trimmed for housing units in 2000; only one of the six was an AIR cluster. For more details on 2000 A.C.E. housing unit weight trimming results, see Chen et al. (2007).

References

Cantwell, P., McGrath, D., Nguyen, N., and Zelenak, M. F. (2001), "Accuracy and Coverage Evaluation: Missing Data Results," DSSD Census 2000 Procedures and Operations Memorandum Series B-7*.

Chen, I., Kilmer, A. and Shores, R. (2007), "Accuracy and Coverage Evaluation: Housing Unit Missing Data Results," DSSD Census 2000 Procedures and Operations Memorandum Series Q-93.

Haines, D. and Mule, T. (2003), "A.C.E. Revision II Results: Change in Estimated Net Undercount," DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-58.

Ikeda, M. and Beaghen, M. (2002), "A.C.E. Revision II – Results from the Noninterview Adjustment," DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-56.

Linse, K. and Stone, C. (2010), "Comparisons and Explanations for Differences in Person Interview Noninterview Rates from the 2000 Accuracy and Coverage Evaluation and the 2010 Census Coverage Measurement Program," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-I-P-02.

Mule, T. (2001), "Accuracy and Coverage Evaluation: Decomposition of Dual System Estimate Components," DSSD Census 2000 Procedures and Operations Memorandum Series B-8*.

Mule, T. (2003), "Summary of Results for Weight Trimming," DSSD Census 2000 Procedures and Operations Memorandum Series #Q-87.

Mule, T. (2008), "2010 Census Coverage Measurement Estimation Methodology," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-18.

Rubin, D.B. (1976), "Inference and Missing Data," Biometrika, 63,581-590.

Shores, R. and Sands, R. (2012), "2010 Census Coverage Measurement Estimation Report: Characteristic Imputation Results," DSSD 2010 Census Coverage Measurement Memorandum Series, 2010-G-7.